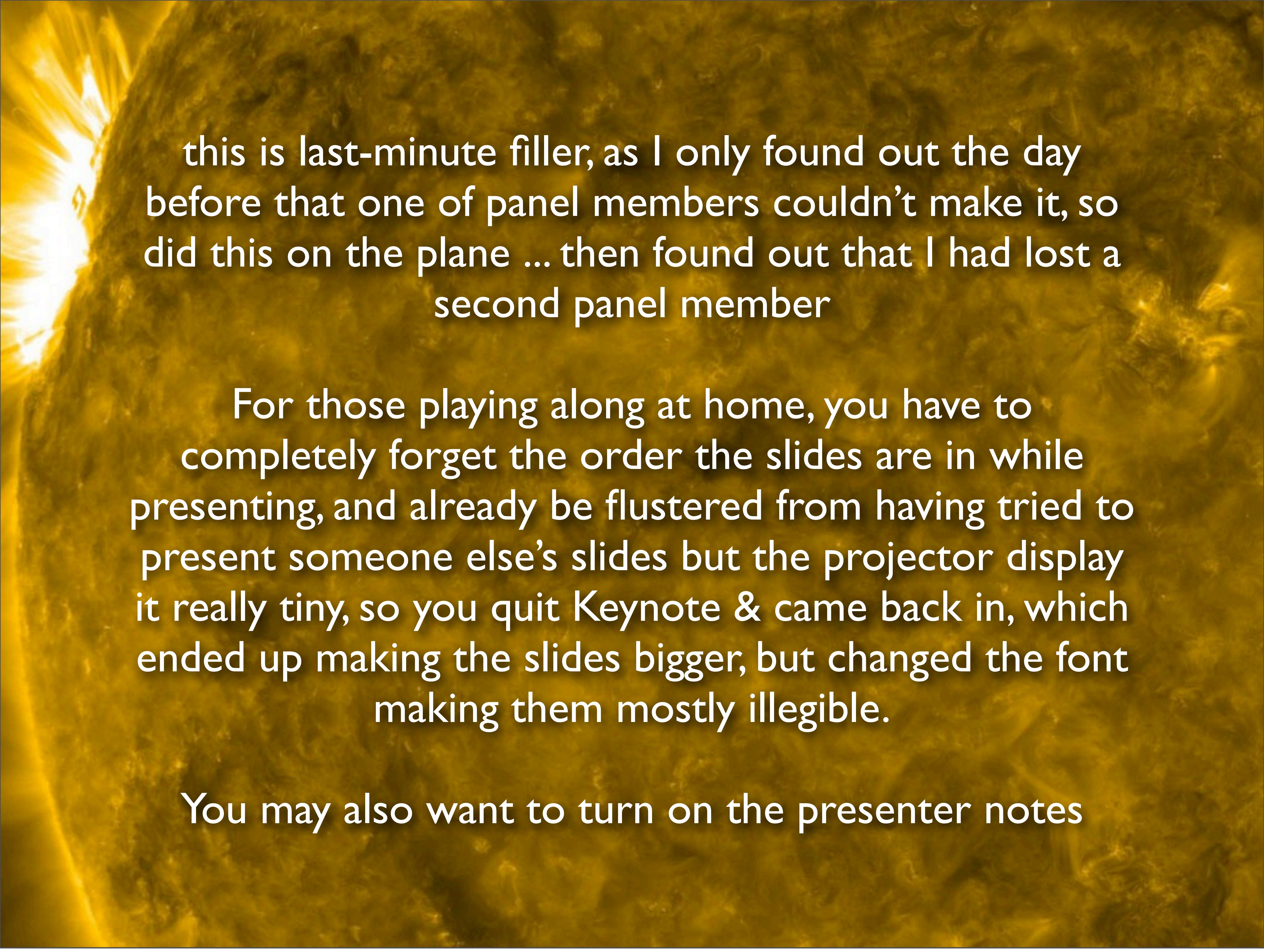
Recommendation “Landing Pages”

RDAP 2012

Yes, I know, I didn’t put my name on the slides ... it was a rush job. (see next slide).
This was presented on 2012-March-22nd, by Joe Hourclé



this is last-minute filler, as I only found out the day before that one of panel members couldn't make it, so did this on the plane ... then found out that I had lost a second panel member

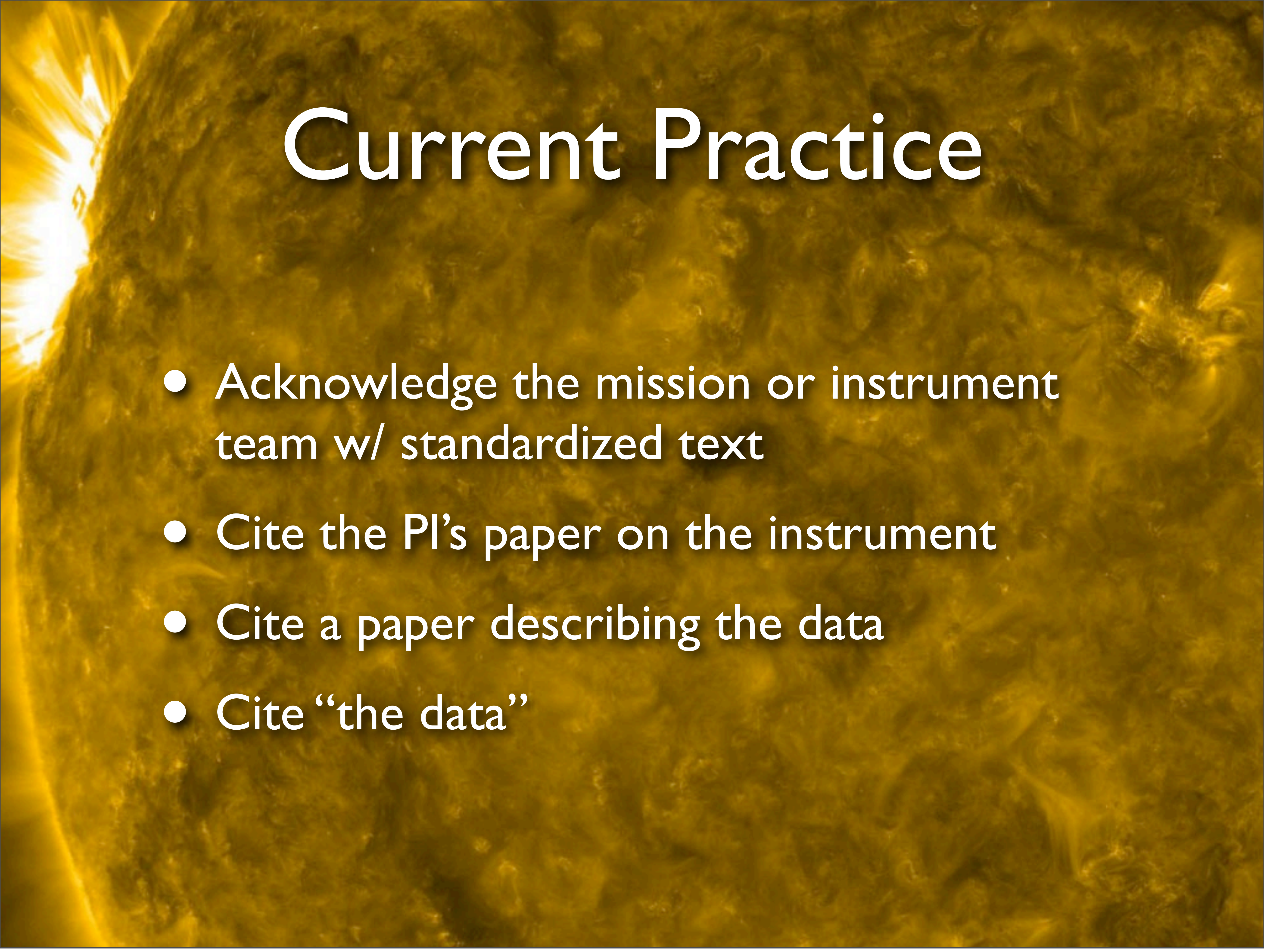
For those playing along at home, you have to completely forget the order the slides are in while presenting, and already be flustered from having tried to present someone else's slides but the projector display it really tiny, so you quit Keynote & came back in, which ended up making the slides bigger, but changed the font making them mostly illegible.

You may also want to turn on the presenter notes

Wednesday, May 2, 2012

This slide was *not* in there when I presented at RDAP.

All of the presenter notes were typed in *after* the presentation, so I've taken the opportunity to try to make things a little clearer than when I presented. I've left those slides untouched, and inserted this slide, and two more at the end.



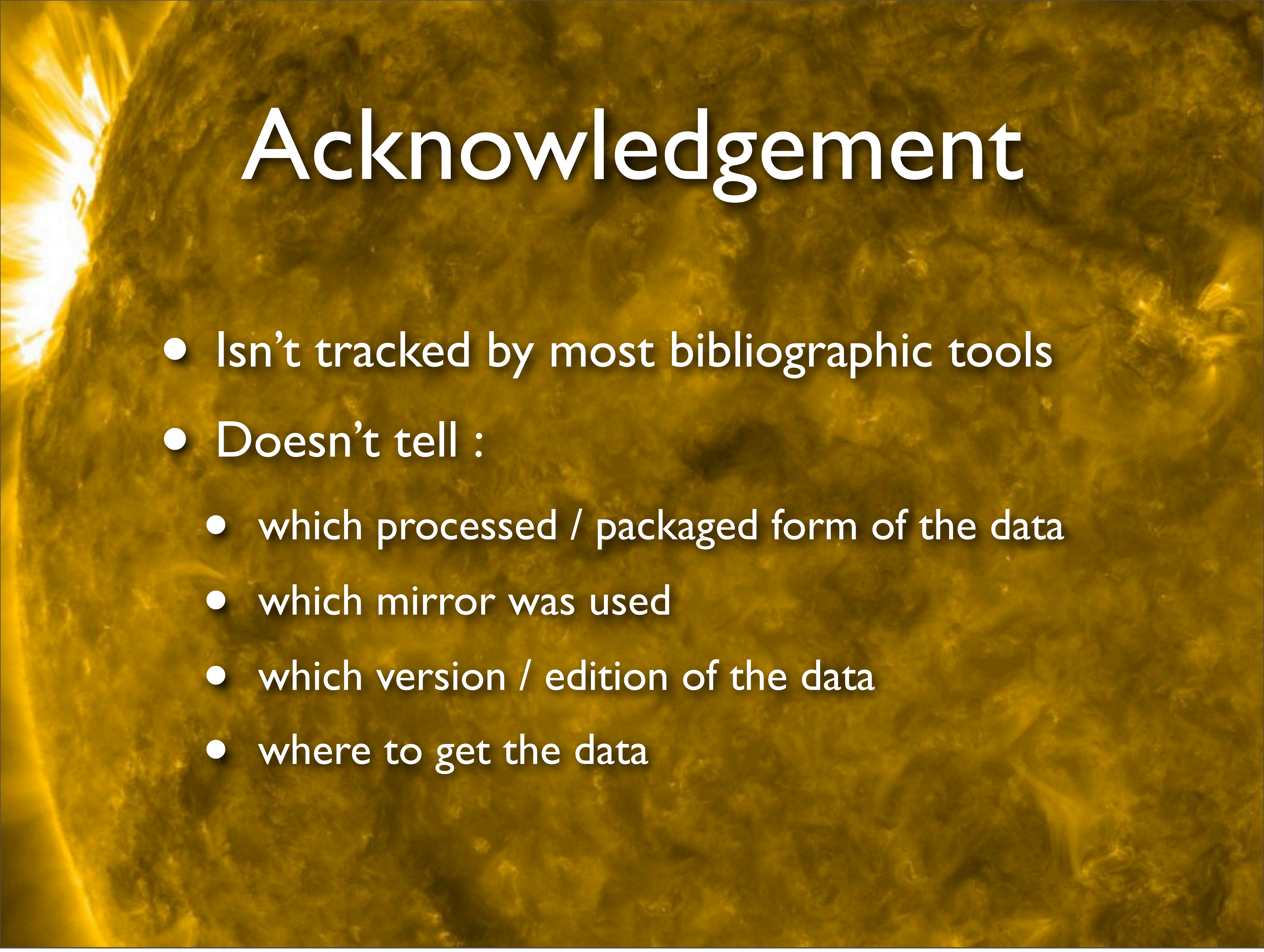
Current Practice

- Acknowledge the mission or instrument team w/ standardized text
- Cite the PI's paper on the instrument
- Cite a paper describing the data
- Cite "the data"

Wednesday, May 2, 2012

And it likely would've helped if I had set the context up correctly ... however, I jumped straight to the material in what's now slide #8, talking about the meeting where these recommendations came from.

Basically, we had a group to try to make a list of the issues of the technical challenges to good data citation. So, first, the quick overview of the current practices in different fields.



Acknowledgement

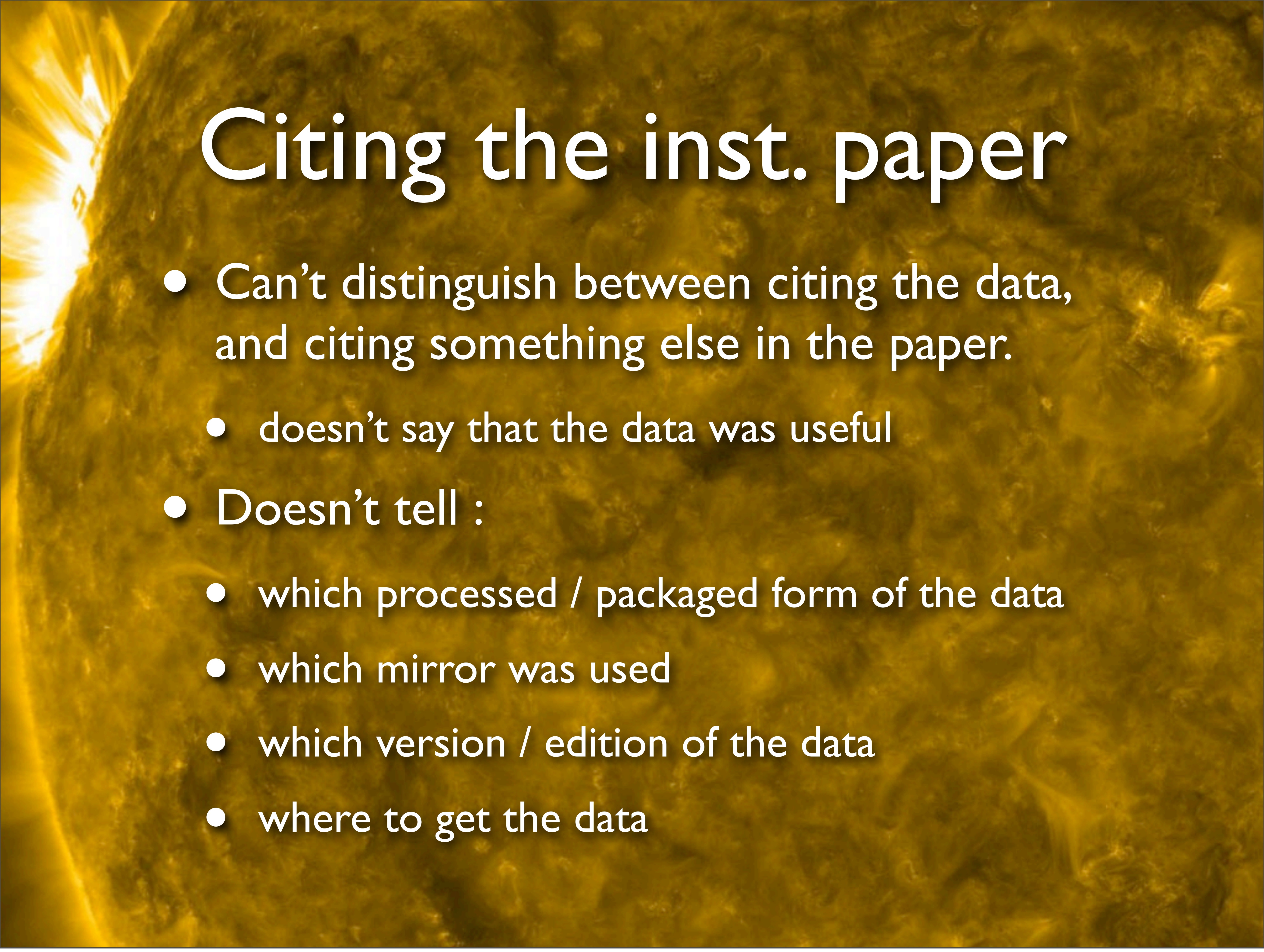
- Isn’t tracked by most bibliographic tools
- Doesn’t tell :
  - which processed / packaged form of the data
  - which mirror was used
  - which version / edition of the data
  - where to get the data

Wednesday, May 2, 2012

An example:

“SOHO is a project of international cooperation between ESA and NASA” ... of course, there’s multiple instruments on the spacecraft, and *lots* of different data products. And it’s been running for 15 years and the data won’t be made ‘final’ until 1-2 years *after* the mission ends (which could be years if they keep getting funding to continue LASCO operations).

And the SOHO ack. text has changed over time ... it used to be that you cited SOHO and the specific instrument (with their PI institutions). But that could mean that you’re mentioning 10+ instruments, so that faded out over time.

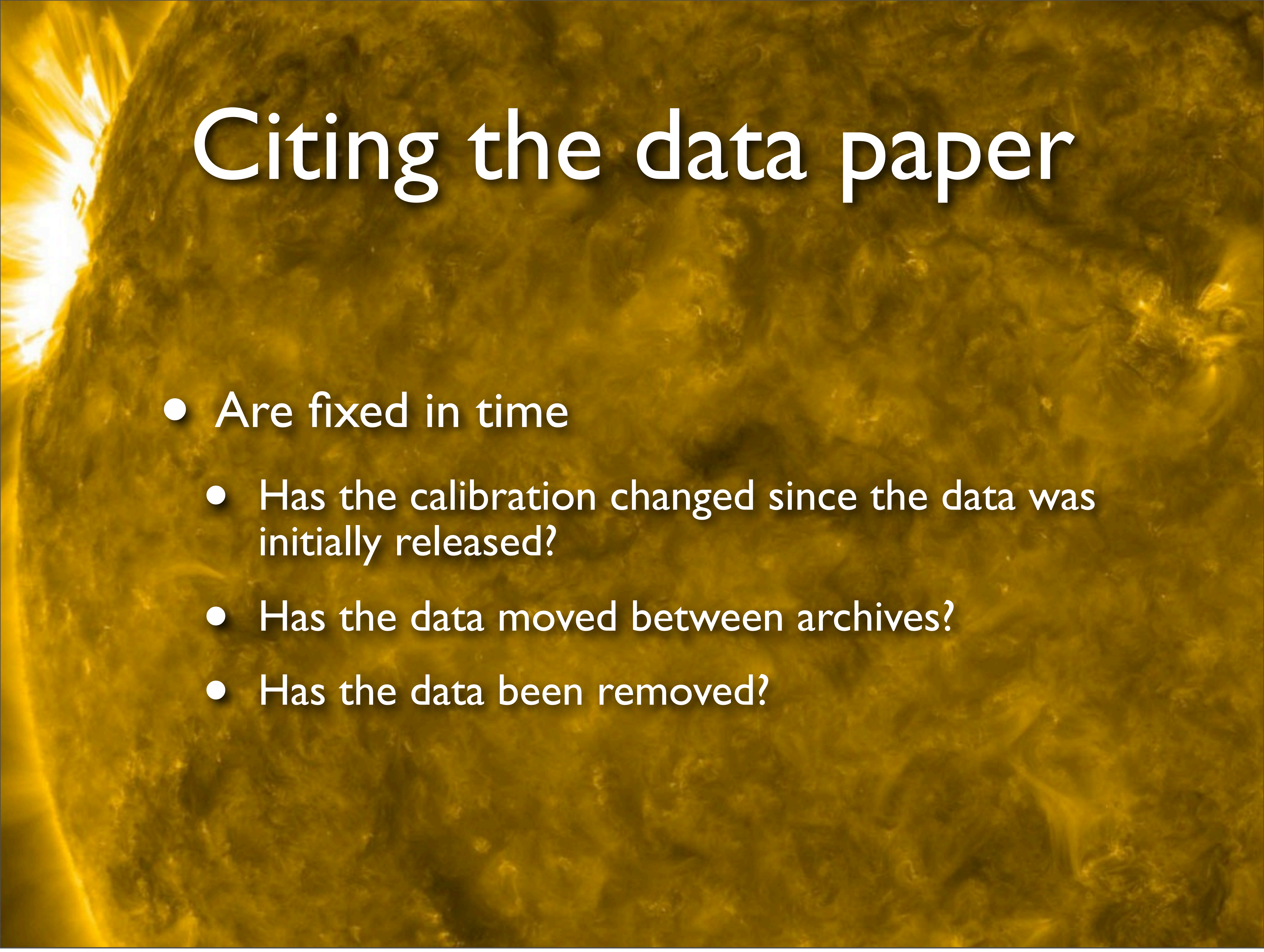


Citing the inst. paper

- Can't distinguish between citing the data, and citing something else in the paper.
  - doesn't say that the data was useful
- Doesn't tell :
  - which processed / packaged form of the data
  - which mirror was used
  - which version / edition of the data
  - where to get the data

Wednesday, May 2, 2012

For instance, you might cite the paper because you're interested in how the mirrors were constructed ... we have no way of differentiating between you're using the data vs. you're interested in the design of the instrument.



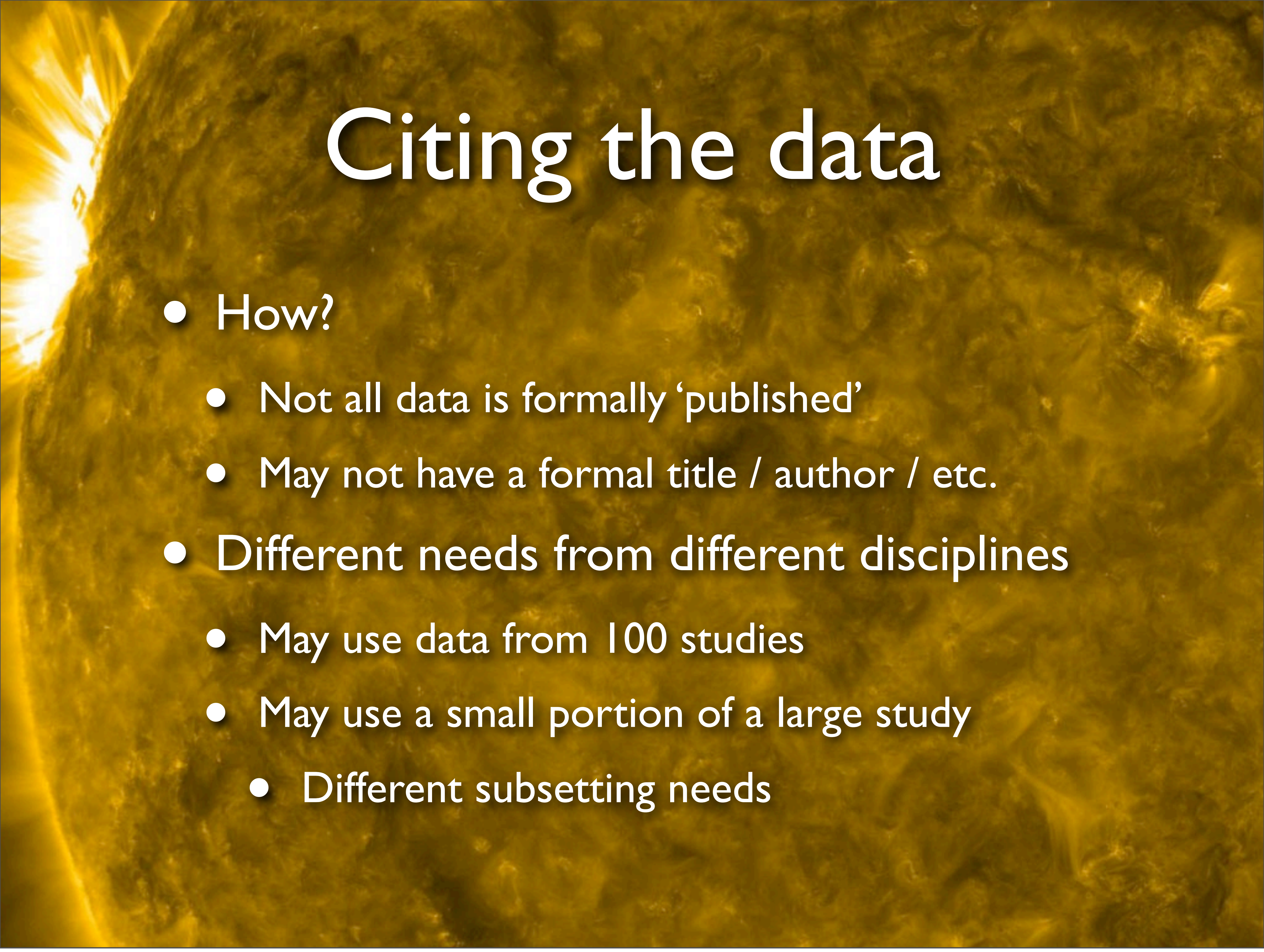
Citing the data paper

- Are fixed in time
  - Has the calibration changed since the data was initially released?
  - Has the data moved between archives?
  - Has the data been removed?

Wednesday, May 2, 2012

The thing about initial calibrations is ... they’re always wrong. They might be pretty close, but there’s always going to be *something* that we learn over time.

For instance, did you know that if you point a telescope at the sun for 15 years, it degrades? Yes, it does. And it wasn’t the PI team who came up with the best model of how the instrument degrades, it was an outside scientist. Should he be the one getting all of the citations whenever someone uses the calibrated data, or the original PI team?



Citing the data

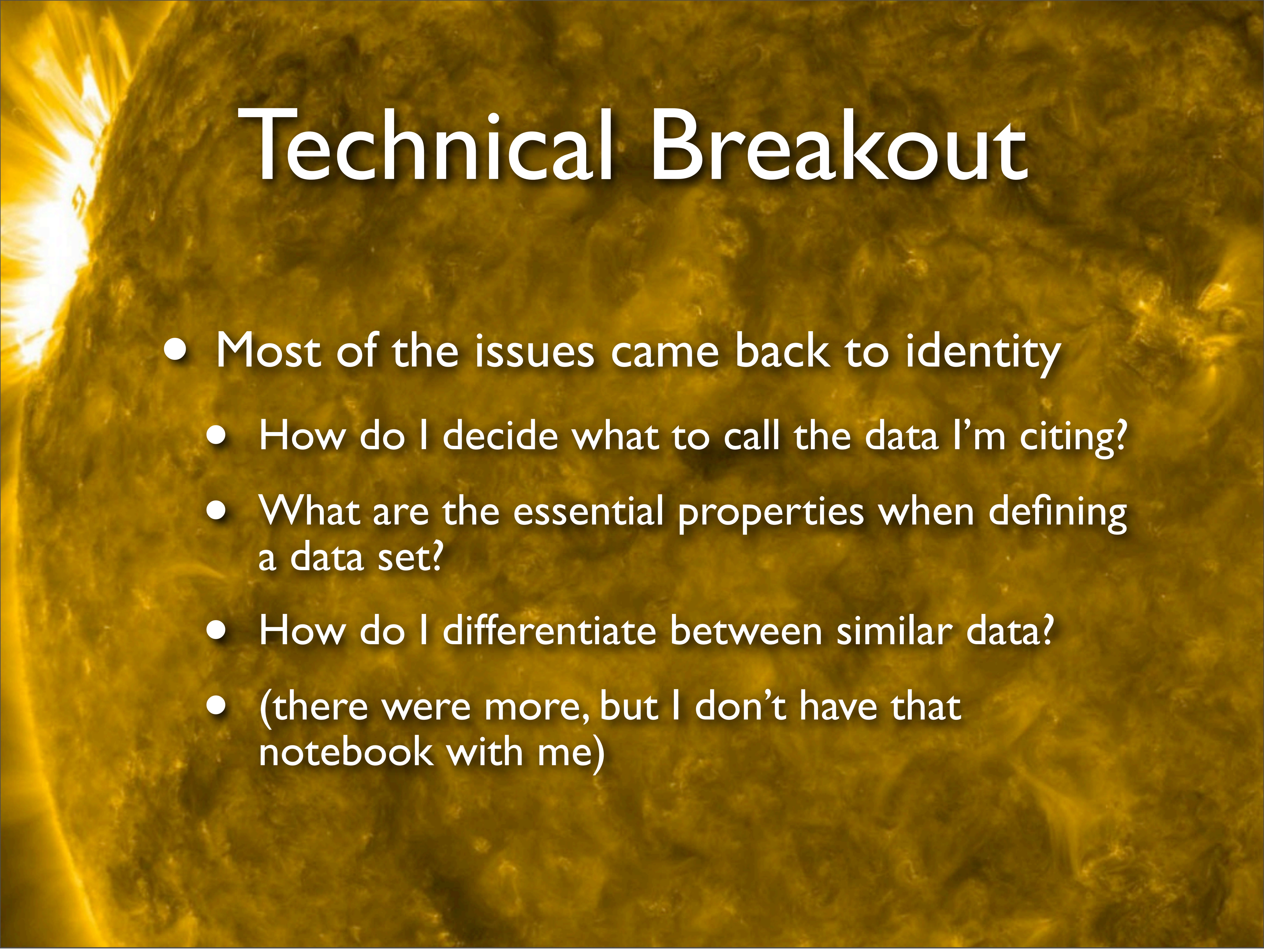
- How?
  - Not all data is formally ‘published’
  - May not have a formal title / author / etc.
- Different needs from different disciplines
  - May use data from 100 studies
  - May use a small portion of a large study
    - Different subsetting needs

Much of the data that is ‘served’ in the solar physics community is a bunch of files dumped on an FTP server or website. They’re FITS files, so they’re self-documenting, but because the instrument might be changing observing modes, it’s hard to define specific ‘collections’ that have titles. Just naming the instrument (or spacecraft, as shown in the SOHO case), is rarely useful to actually

BRDI Meeting, Aug. 2011

- Two days discussing data attribution & citation
- Breakouts made lists of the challenges & issues concerning attribution & citation
  - Technical, Scientific, Institutional/Financial/Legal/ Socio-Cultural, Main Actors & Roles, Add’l info needed to proceed.

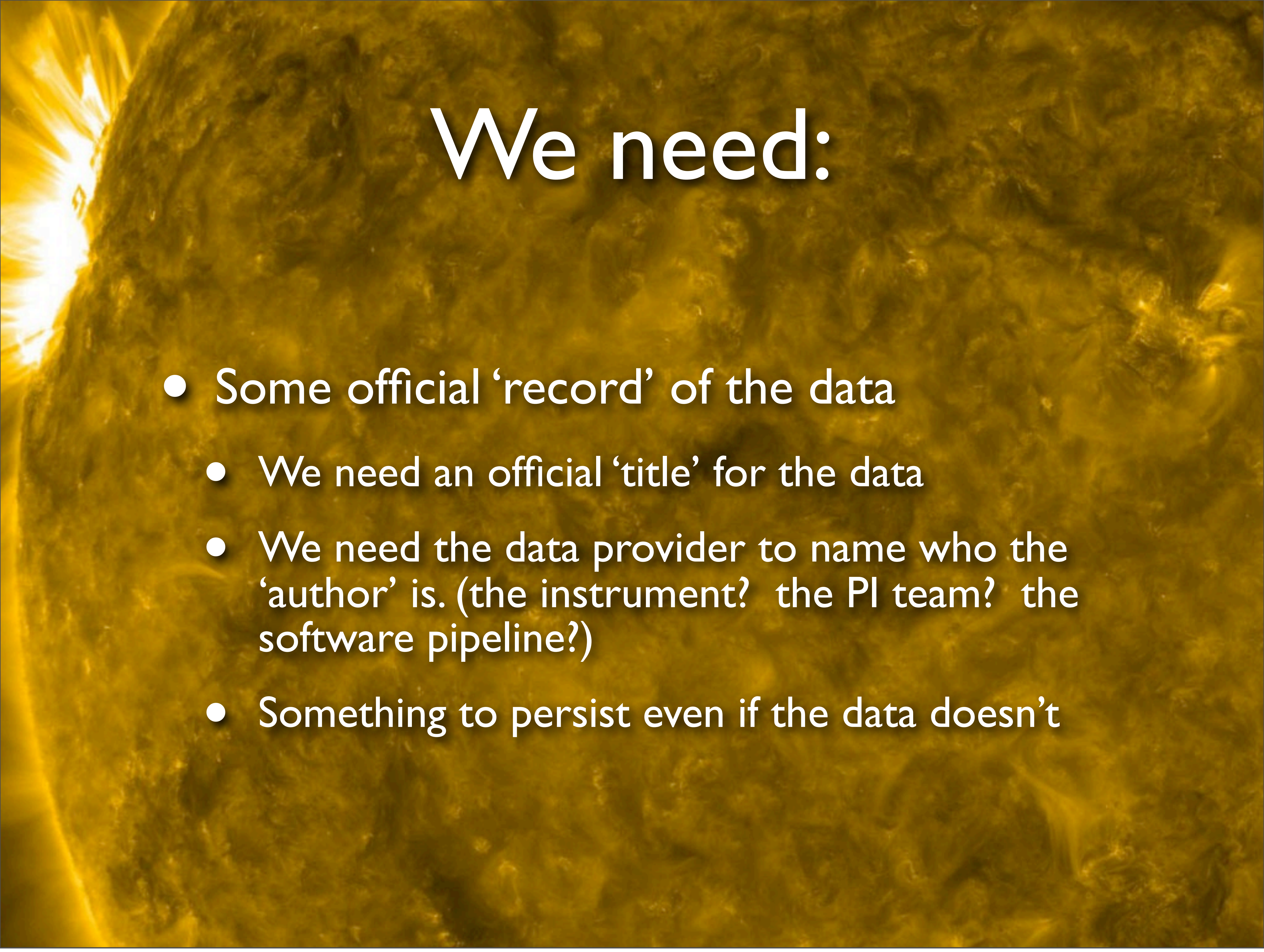
BRDI meeting program w/ links to presentations is at:
http://sites.nationalacademies.org/PGA/brdi/PGA_064019



Technical Breakout

- Most of the issues came back to identity
  - How do I decide what to call the data I’m citing?
  - What are the essential properties when defining a data set?
  - How do I differentiate between similar data?
  - (there were more, but I don’t have that notebook with me)

I should’ve mentioned that we were trying to look at the technical barriers to citation. Part of the issue was that we weren’t sure if it was the researchers or the data provider’s responsibility on generating the citations.

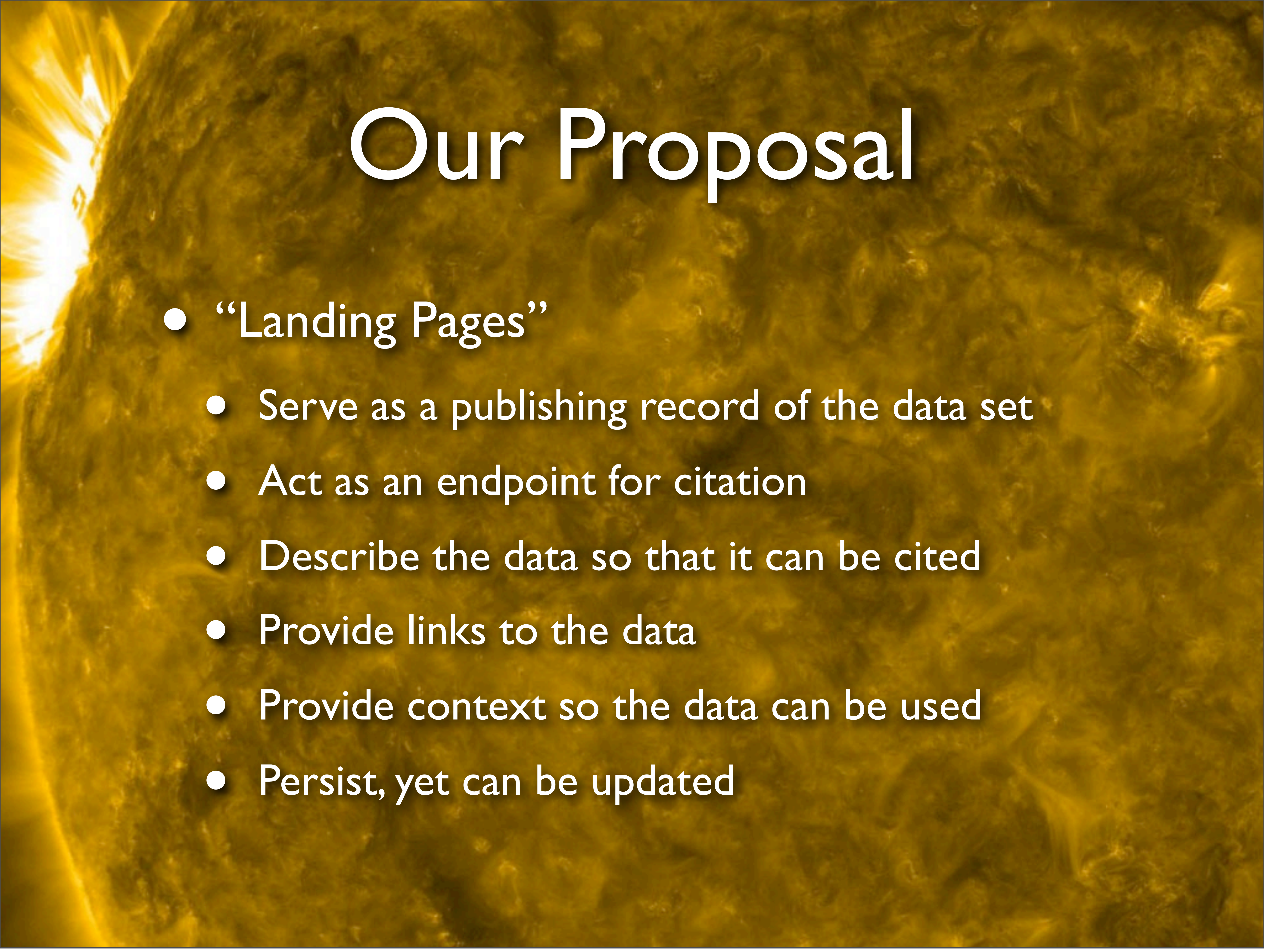


We need:

- Some official ‘record’ of the data
  - We need an official ‘title’ for the data
  - We need the data provider to name who the ‘author’ is. (the instrument? the PI team? the software pipeline?)
  - Something to persist even if the data doesn’t

Wednesday, May 2, 2012

Basically, if you want people to cite your data in a specific way, you need to tell them how to cite it. Don’t let them pick the ‘author’, or the ‘title’, as there could be infinite variation. If you want to glob lots of stuff under one title (a ‘collected works of Shakespeare’, you can do it, but it’s not nearly as useful as being able to cite one of the individual plays.)



Our Proposal

- "Landing Pages"
  - Serve as a publishing record of the data set
  - Act as an endpoint for citation
  - Describe the data so that it can be cited
  - Provide links to the data
  - Provide context so the data can be used
  - Persist, yet can be updated

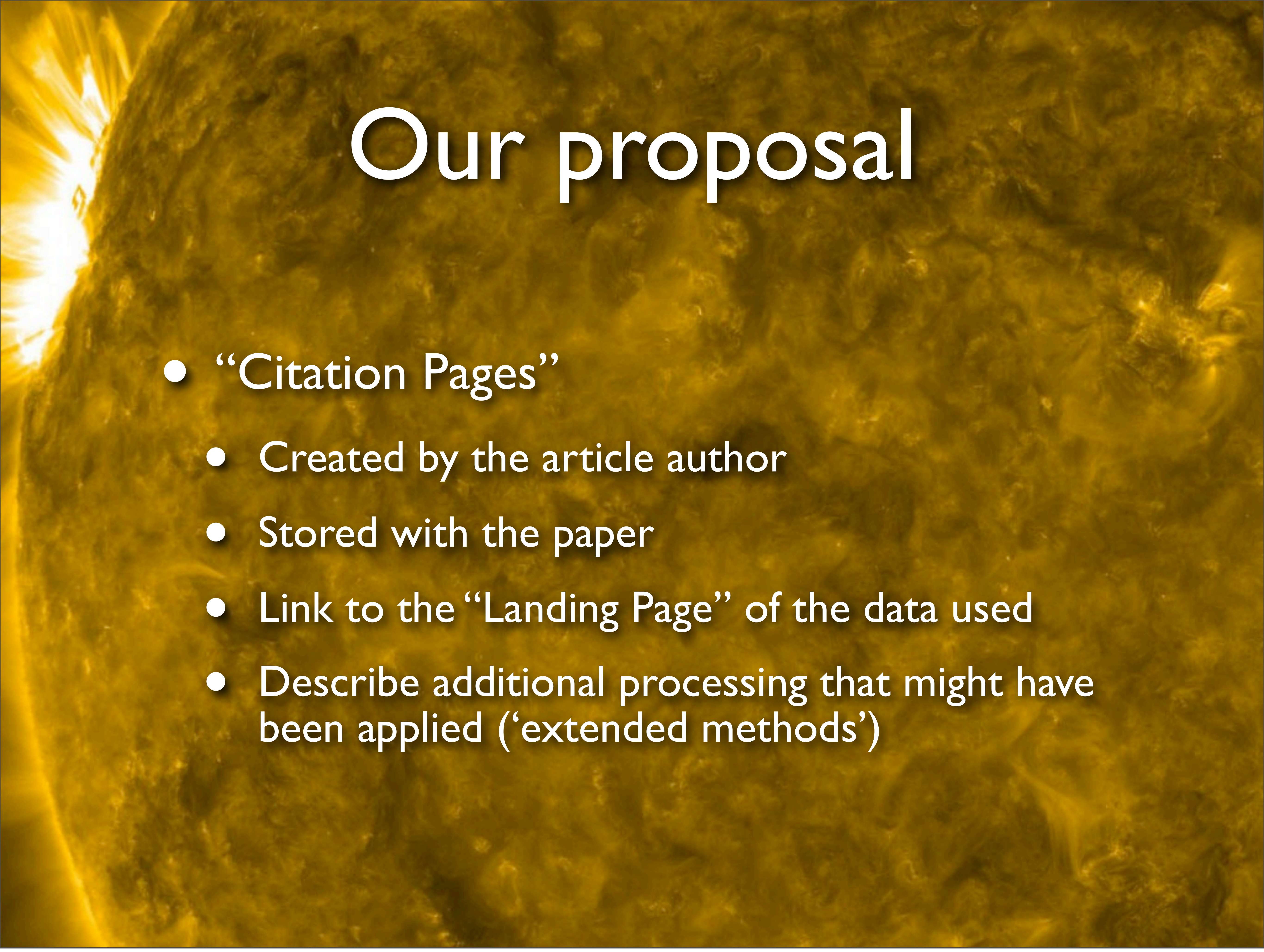
Leaving the citation generation up to the individual researcher won't work ... but also, giving a specific 'you must use this format' isn't as useful, because the author may not be using APA or whatever style guide your discipline's journal uses. (this is where DataCite is useful)

There's no reason why we can't describe the data *once* using DataCite, DublinCore-SAM (once that profile is done) or similar, and present it in BibTex or other formats for citation manager software to use. The search engines could even spit out a little report when you download data, as a sort of 'receipt', telling you what you've just downloaded, and how to cite it.

We also don't specifically define the granularity of the collection described -- it could be an individual file; all data for the life of the investigation; broken out by year; divided up by observing mode ... or you could do multiple, as there's nothing to say that a file can't be a member of more than one collection; the provider can define whatever groupings make sense for their data.

The landing pages should also be machine harvestable -- XML w/ XSLT transforms to make it XHTML, RDFa, HTML+microformats, HTTP content negotiation, links to OAI-ORE links, etc. If you have lots (but not TBs) of files to download, it can link to OAI-ORE, MetaLink, sparse bags (BagIt) or similar to automate downloading to those files.

... I could keep typing, but it might make more sense to just go read the handout. (see the last slide for a link)



Our proposal

- “Citation Pages”
  - Created by the article author
  - Stored with the paper
  - Link to the “Landing Page” of the data used
  - Describe additional processing that might have been applied (‘extended methods’)

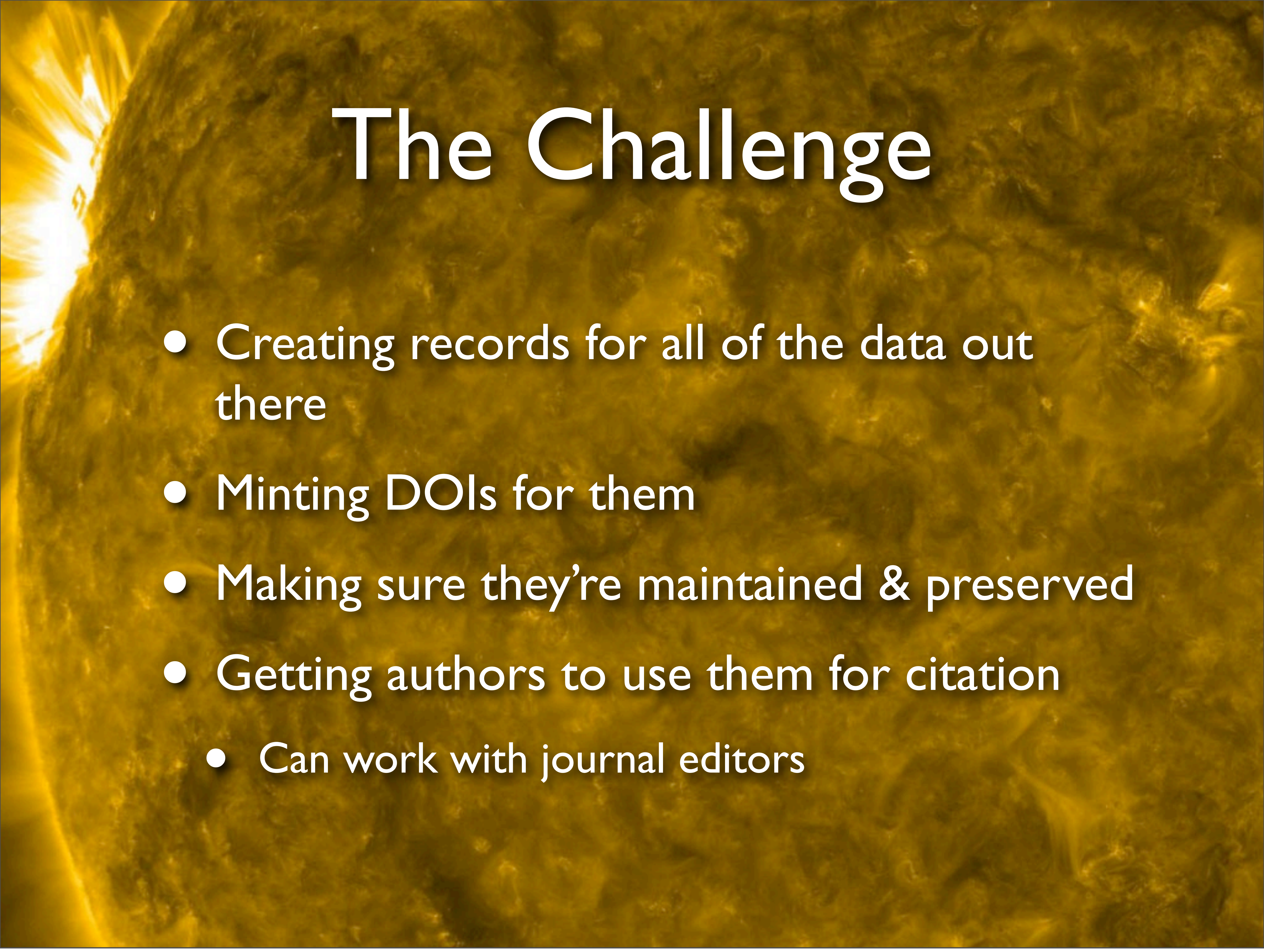
Wednesday, May 2, 2012

If the journal publisher won’t take it, then in an Institutional Repository, or even on the author’s website is better than nothing.

As the data provenance models mature (eg, the work at globalchange.gov), we may be able to make the processing description machine readable.

In those cases where you can store the data with the publication (eg, the data’s small enough), you can do it, but you still want to describe how it came to be for reproducability. The citation page can also be used for ‘enhanced publications’ or ‘interactive data pages’, to allow readers of the paper to quickly visualize the data, change the scaling or field of view of a plot; reorder or filter a tabular data, etc.

The citation page may just link to a single published dataset, or it may aggregate hundreds of datasets. If you only have a few (<5) sources of data, consider citing them individually rather than aggregated, even if you have limits as to how many citations you’re allowed, as without the data, you wouldn’t have been able to do the research.

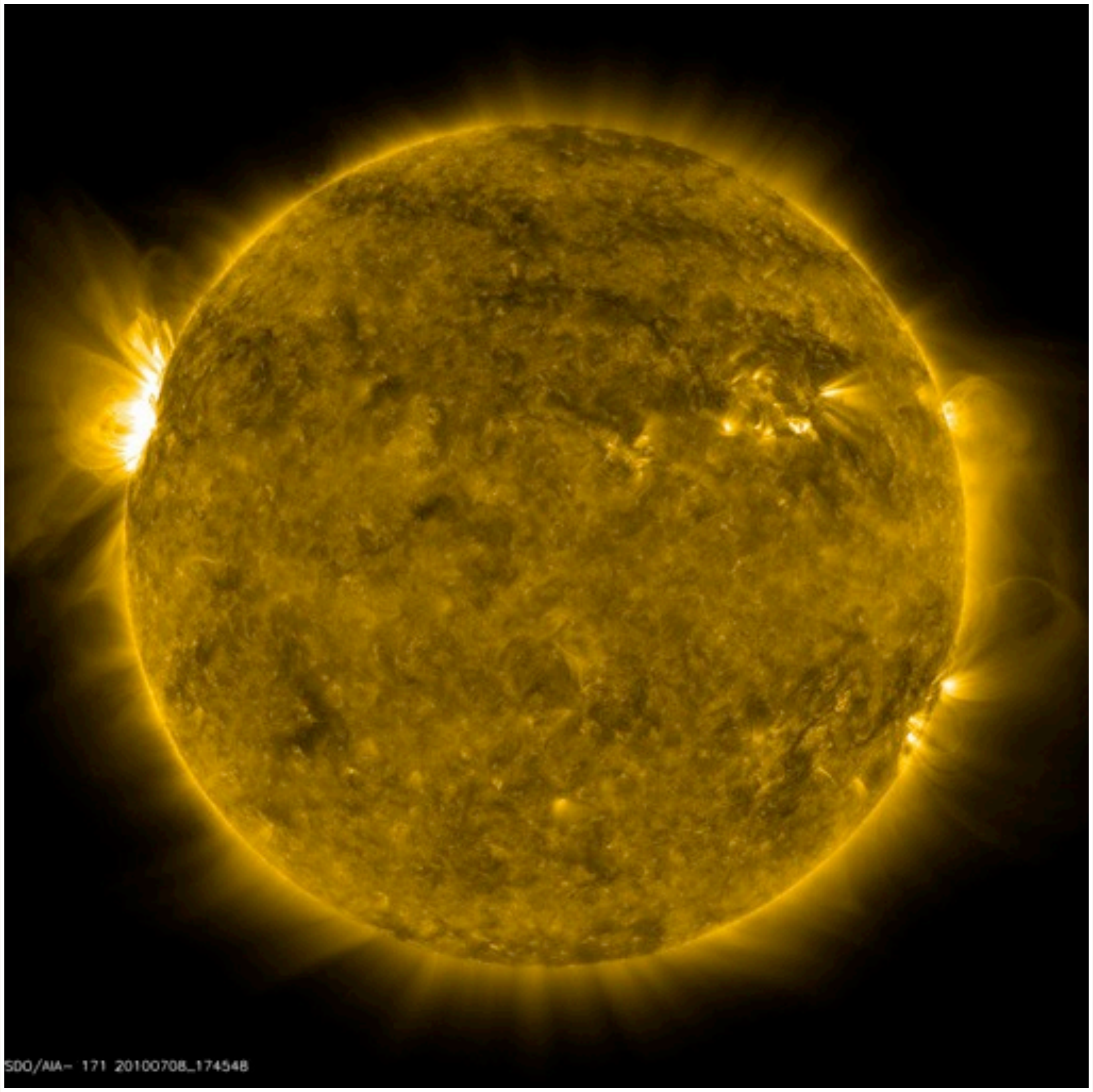


The Challenge

- Creating records for all of the data out there
- Minting DOIs for them
- Making sure they're maintained & preserved
- Getting authors to use them for citation
  - Can work with journal editors

Wednesday, May 2, 2012

If we're going to do this efficiently, we likely need to do it by discipline -- each data archive describing the data would be nice, but it's a lot of time reading documentation, interpreting what each field/attribute means for their data, etc. We get an efficiency of scale working by discipline. We'd likely need at least two people -- (1) a person experienced with metadata / cataloging, and (2) a discipline scientist who can verify the work, and possibly write missing documentation. New post docs often write the best documentation, as they come at it fresh, having to research things, rather than just assuming 'everyone' knows about that little quirk that's not worth mentioning to someone in the field, but would likely not be known by people from other disciplines.



Solar Dynamics Observatory (SDO)
Atmospheric Imaging Assembly (AIA)
171Ångstrom ; 2010/07/08 17:45:48UT ; 2x2 binned

This slide was *not* in there when I presented at RDAP.

I usually acknowledge the image I’m using in the background ... even had a master slide so I could insert it easily ... but I forgot to put it in. And this time, I’ve linked it to the mocked up ‘landing page’ for SDO/AIA Lev1 171 Angstrom images.

(which I should probably go and modify slightly, to explain that it’s a mockup, now that I’m giving out the URL to it)

Poster, Handout & links:

http://docs.virtualsolar.org/wiki/Citation

Wednesday, May 2, 2012

This slide was *not* in there when I presented at RDAP.

The link goes to the Virtual Solar Observatory's wiki. It has links to all of the references mentioned in the poster & handouts ... and I should probably add 'MetaLink' and 'BagIt' and other stuff.